Design for Standard Setting: Arizona 2008

CTB/McGraw-Hill

May 2, 2008

Background

In 2008, the Arizona Department of Education (ADE) will engage in a multi-step process to establish cut scores for the Arizona Instrument to Measure Standards (AIMS) Science for Grades 4, 8, and high school. The multi-step process will be comprised of educator cut score recommendations, recommendations for smoothing educator-suggested cut scores, and final approval by the State Board of Education. The ADE has indicated that cut scores will be established for *Exceeds the Standard*, *Meets the Standard*, *Approaches the Standard*, and *Falls Far Below the Standard*.

The AIMS Science will be administered operationally for the first time in the Spring 2008. The AIMS Science will be administered to students in Grades 4 and 8 and will be administered to Grade 10 students enrolled in a life science course. Grade 9 students enrolled in a life science course may opt to take the AIMS Science in Grade 9. The AIMS Science for high school is not required for high school graduation. (See http://www.ade.state.az.us/standards/downloads/AZAssessmentOverview-Aug07.ppt#256,1,Arizona's Assessment Program.)

The standard setting workshop is tentatively scheduled to occur on June 9 – 11, 2008 for the AIMS Science in Grades 4, 8, and high school. Table 1 shows the day-by-day timelines for the standard setting workshops.

This document describes the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) which will be used to allow Arizona educators to recommend cut scores. This document also describes the smoothing process that immediately follows the BSSP. It does not describe the process that the State Board of Education will use to give final approval to cut scores.

Included in Appendix A of this document is the *Bookmark Standard Setting Handbook* (2005). This Handbook outlines the typical implementation of the Bookmark Standard Setting Procedure. It also includes generic versions of the training materials and overheads that CTB uses to train participants. This Handbook also provides guidelines to sponsoring agencies (ADE) on how to recruit participants.

Table 1. Day-by-day timelines for the standard setting to be held in 2008.

Day	Time	Activity
Day 1	**AM**	Table Leader Training
	PM	Opening Session
Day 2	**AM**	Bookmark Activities
	PM	Bookmark Activities
Day 3	**AM**	Bookmark Activities
	PM	Smoothing Activity

Bookmark Standard Setting Procedure

The BSSP is typically implemented using item-response theory (IRT); however, it may be implemented using classical test theory when the test does not meet the conditions for using IRT, such as the case during the 2005 AIMS standard setting for Reading and Mathematics (see *Arizona's Instrument to Measure Standards Bookmark Standard Setting Technical Report (2005*) for a complete description of the BSSP). Primarily, the outcome of the standard setting workshop should not be predetermined or unreasonably limited at the outset of the standard setting by the standard setting method employed.

This section provides data analysis in support of using an IRT approach for the BSSP. Furthermore, the analyses provide guidance for the Arizona National Assessment and Accountability Advisory Committee (NAAAC) regarding which response probability (RP) criterion should be used for the 2008 implementation of the BSSP.

In the BSSP, items are presented to panelists along the reporting scale continuum from the easiest item to the most difficult item. Typically, items are mapped to the scale based upon the response probability .67 (RP67). The RP criterion is the probability that a student with a particular scale score will answer a given item correctly. In the Rasch model, the scale score is linearly related to the item difficulty and can easily be solved given the RP and where *b* is the item difficulty parameter.

$$\theta = b + \ln \frac{RP}{1 - RP}$$

RP67 indicates that items are mapped to the scale at the point where a student with a score at the mapping location would have a 67% chance of answering the item correctly.

Because intact science forms have not been operationally administered, CTB projected 2008 student performance using field test data. A description of the field test data, procedures utilized, and a summary and discussion of the results follow.

Field Test Data

In 2007, the ADE field tested five forms of science items in Grade 4, Grade 8, and high school. Each field test form was composed of 42 items. The field test forms were spiraled within classrooms with approximately 2,000 students taking each form as shown in Table 2.

Table 2. Field test sample

Grade	Form	N
4	1	2347
4	2	2323
4	3	2334
4	4	2322
4	5	2219
8	1	2383
8	2	2359
8	3	2364
8	4	2257
8	5	2223
HS	1	1886
HS	2	1862
HS	3	1810
HS	4	1793
HS	5	1748

The field test data were used to build operational tests[1]. The operational tests comprise a 54-item Grade 4 Science test, a 58-item Grade 8 Science test, and a 65 item high school Science test. Two operational forms of each test (A and B) were constructed for each grade, neither of which has been administered intact operationally to date. In order to investigate projected student performance in 2008, CTB equated the field test data across forms using the following procedures.

Projected Data Procedures

Parameters for each item were obtained during the field test administration. Each field-test form was independently calibrated using the Rasch model. Because forms were spiraled within classrooms, the random groups design (Kolen & Brennan, 2004) could be used to equate the five forms onto the same scale. Kolen and Brennan (2004) noted that when using this design, group-level performance differences are taken as an indication of the difference in difficulty among the forms (p. 15).

To equate the items across forms, Form 1 was used as the baseline to set the logit scale. Items were then placed on the same scale using mean sigma equating. Because a scale has not yet been set for science, a temporary scale was produced using the scaling transformation constants from the same grade AIMS mathematic assessments. This permitted a raw score to scale score table to

[1] The only operational forms that will exist are the two established by the census test in spring 2008.

be constructed that allowed for student scores to be estimated on the same scale for each of the five forms.

Next, using the equated parameters for items used to build forms A and B, a raw score to scale score table was built for Form A and Form B in each grade. The AIMS Math lowest obtainable scale score (LOSS) was assigned to the 0 raw score point and any other raw score points for which the scale score was below the LOSS. Similarly, the AIMS Math highest obtainable scale score (HOSS) was assigned to the maximum raw score point and any other raw score points for which the scale score was above the HOSS. The HOSS and LOSS were evaluated to determine whether adjustment was necessary. No adjustments were made.

Tables for Form A and Form B were compared. The two tables were similar which was to be expected because items for the forms were selected based on similar test characteristic curves for the two forms using field test parameters.

To yield projected scale scores for students on Form A and Form B, the student scale scores on each of the 5 field test forms had to be mapped onto Form A and Form B. Each student was assigned the closest scale score in the scoring table for Form A or Form B based on the scale score estimate for the student based upon the field test form data. The item location and projected student data on each form were then used to investigate the feasibility of using an IRT-based Bookmark procedure.

Results

In an ideal distribution of items for the purposes of standard setting, items would be evenly distributed across the 10 deciles. In the 2001 text *Setting Performance Standards*, Mitzel, Lewis, Patz, and Green noted:

> Ordered item booklets span from about 80 to 110 score points, which exceeds normal test lengths. We view the ability to present a more representative sample of a content domain than a single test form to be a strength of the procedure (p. 252).

This means optimally, between 8 and 11 items are present for each decile that describe what students know and are able to do.

If the ADE determines that panels of participants will set standards on an intact test form, 5.4 items would be found in each decile range for the Grade 4 test, 5.8 items would be found on the Grade 8 Science test, and 6.5 items would be found in each decile range on the HS test. Tables 2−4 compare the results of the projected data for Form A to the ideal by showing the difference between the projected number of items at or below each percentile in Form A to what would be optimal for several RP values. Because Form A and Form B are similar, the review of the data shown below is for Form A only. Positive numbers indicate more items are present than the ideal to describe student abilities and negative numbers indicate fewer items are present than the ideal to describe student abilities.

For example, consider that Form A of the Grade 4 Science test has 54 items. Ideally, the first decile would include one-tenth of these items, or 5.4 items. Similarly, each subsequent decile

would ideally contain an additional 5.4 items. As shown in Table 2, the large positive difference of 13.2 found with at decile 20 with RP50 suggests that there are more than 13 additional items at this decile than the ideal. This indicates that there are more items towards the lower end of the scale than would be expected, and as a result, there will be fewer items in the upper portion of the distribution.

To approach the ideal distribution of items, the difference between the number of items at or below selected percentiles, as illustrated in Tables 3, 4, and 5, should be near zero. Larger numbers, whether positive or negative, indicate that there are too few items at some given point within the distribution. For example, differences greater than six can be found with RP50 and RP55 in Grade 4 which indicate that there are more items towards the lower end of the scale than would be expected, and as a result, there will be fewer items in the upper portion of the distribution.

Table 3. Difference between number of items at or below selected percentiles and the ideal number of items for Grade 4 Science: Form A.

Grade 4 Science						
Decile	**Ideal**	**RP50 Difference**	**RP55 Difference**	**RP60 Difference**	**RP65 Difference**	**RP67 Difference**
10	5.4	5.6	3.6	0.6	0.6	0.6
20	10.8	13.2	10.2	8.2	3.2	1.2
30	16.2	11.8	9.8	7.8	5.8	4.8
40	21.6	8.4	6.4	6.4	4.4	4.4
50	27	8.0	8	3	1.0	1
60	32.4	7.6	6.6	2.6	2.6	1.6
70	37.8	8.2	5.2	2.2	0.2	-2.8
80	43.2	9.8	4.8	4.8	2.8	2.8
90	48.6	5.4	5.4	4.4	4.4	4.4
100	54	0.0	0.0	0.0	0.0	0
Mean Difference		7.8	6	4	2.5	1.8
Standard Deviation		3.7	3.0	2.8	2.0	2.4

Table 4. Difference between number of items at or below selected percentiles and the ideal number of items for Grade 8 Science: Form A.

Grade 8 Science						
Decile	**Ideal**	**RP50 Difference**	**RP55 Difference**	**RP60 Difference**	**RP65 Difference**	**RP67 Difference**
10	5.8	2.2	1.2	0.2	-0.8	-0.8
20	11.6	5.4	0.4	-0.6	-3.6	-3.6
30	17.4	4.6	0.6	-1.4	-5.4	-6.4
40	23.2	8.8	4.8	-1.2	-5.2	-5.2
50	29	9	6	3	-3	-6
60	34.8	10.2	5.2	0.2	-2.8	-2.8
70	40.6	11.4	10.4	4.4	-0.6	-2.6
80	46.4	7.6	6.6	5.6	4.6	0.6
90	52.2	2.8	1.8	1.8	1.8	0.8
100	58	0	0	0	0	0
Mean Difference		6.2	3.7	1.2	-1.5	-2.6
Standard Deviation		3.8	3.4	2.4	3.1	2.7

Table 5. Difference between number of items at or below selected percentiles and the ideal number of items for HS Science: Form A.

HS Science						
Decile	**Ideal**	**RP50 Difference**	**RP55 Difference**	**RP60 Difference**	**RP65 Difference**	**RP67 Difference**
10	6.5	1.5	0.5	-0.5	-1.5	-2.5
20	13	4	1	-3	-5	-6
30	19.5	3.5	2.5	-1.5	-3.5	-6.5
40	26	6	1	-3	-4	-6
50	32.5	3.5	2.5	-1.5	-5.5	-5.5
60	39	5	2	-3	-5	-6
70	45.5	4.5	2.5	0.5	-4.5	-4.5
80	52	5	0	-2	-4	-5
90	58.5	4.5	4.5	3.5	0.5	-1.5
100	65	0	0	0	0	0
Mean Difference		3.75	1.65	-1.05	-3.25	-4.35
Standard Deviation		1.8	1.4	2.0	2.2	2.2

It should be noted that Tables 3–5 show summative values at specific distribution points, and more detailed and critical information may be found by viewing the distribution of items graphically. The tables will hide items that cluster within certain deciles. Figures 1−3 show the

percent of students at or below each page in the ordered item booklet (OIB) using RP67. Through a careful review of the figures several technical issues of note may be viewed.

First, in several areas in each test form, items have the same bookmark location (e.g., the first few items in Grade 4 Form A). In the Rasch model, this means that these items have approximately the same difficulty parameters though they may measure different areas of the test blueprint. From an IRT perspective, however, these items are providing the same information about student abilities. Though this is not uncommon in test forms, if a standard setting panel is placing a bookmark in areas where this occurs, they may become frustrated; when participants move a bookmark up or down in an OIB they expect the content changes underlying an item to represent cut score changes as well. When OIBs are longer than the intact test this issue may be solved by selecting items to represent the test blueprint while minimizing the number of items that have the same bookmark location.

When intact test forms have several items that are at the same location and the intact test form is shorter than the typical OIB, one will often see gaps in the student distribution such that a change in one page of an OIB means a change in impact data of 3%−6% (e.g., Grade 4 Form A between items 35 and 36). This means that the placement of the cut scores along the theta distribution may be less precise than might be desired. Given that the impact data in Figures 1–3 is for the total field test population, these gaps will be wider for the field test subpopulations. As with the previous issue, longer OIBs oftentimes solve this because more items are present to describe the theta distribution and so fewer gaps will often be observed.

Figure 1. Projected Impact Data using RP67 For Grade 4 Form A.

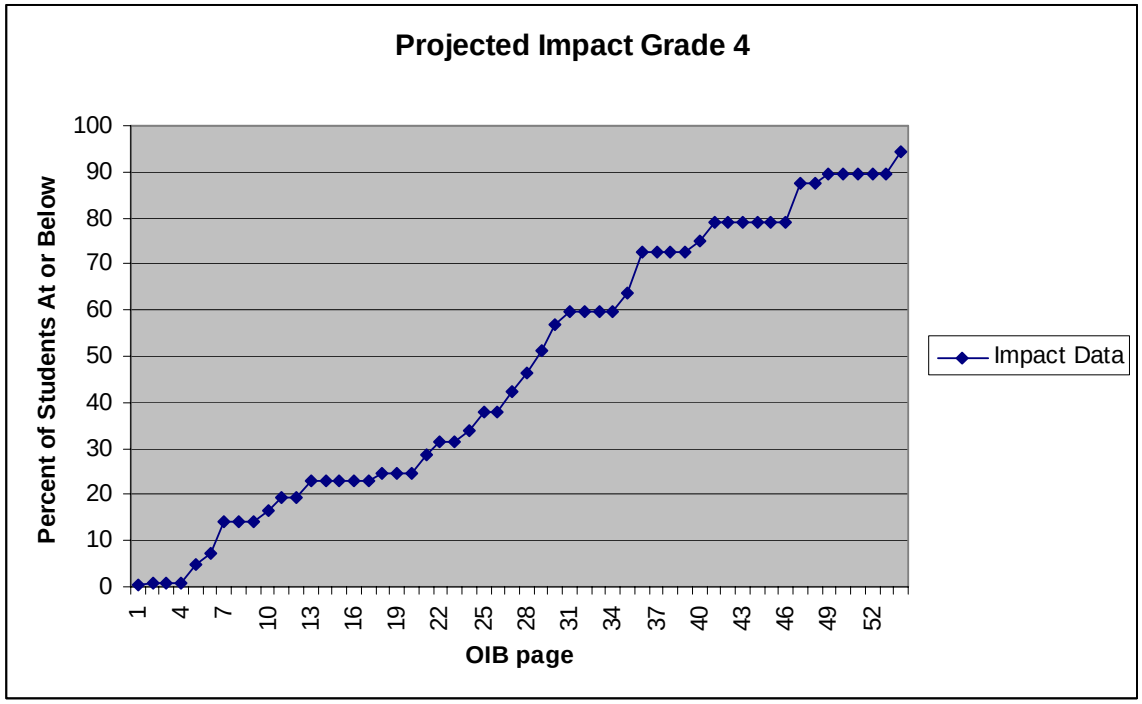


Figure 2. Projected Impact Data using RP67 for Grade 8 Form A.

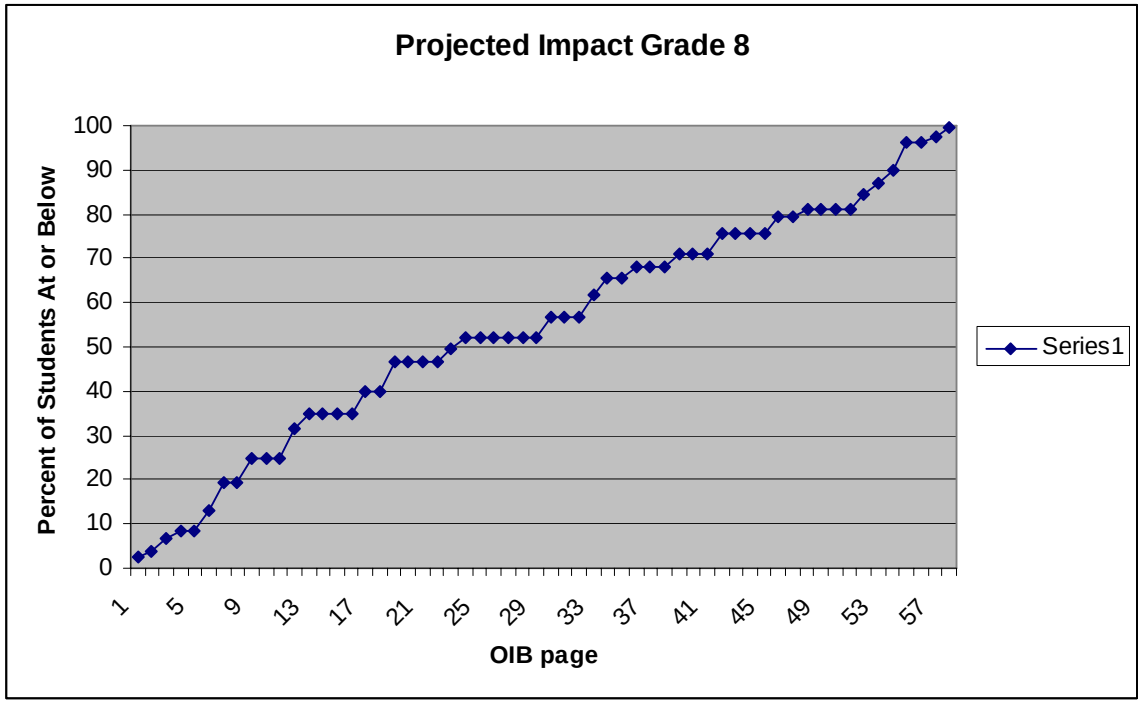
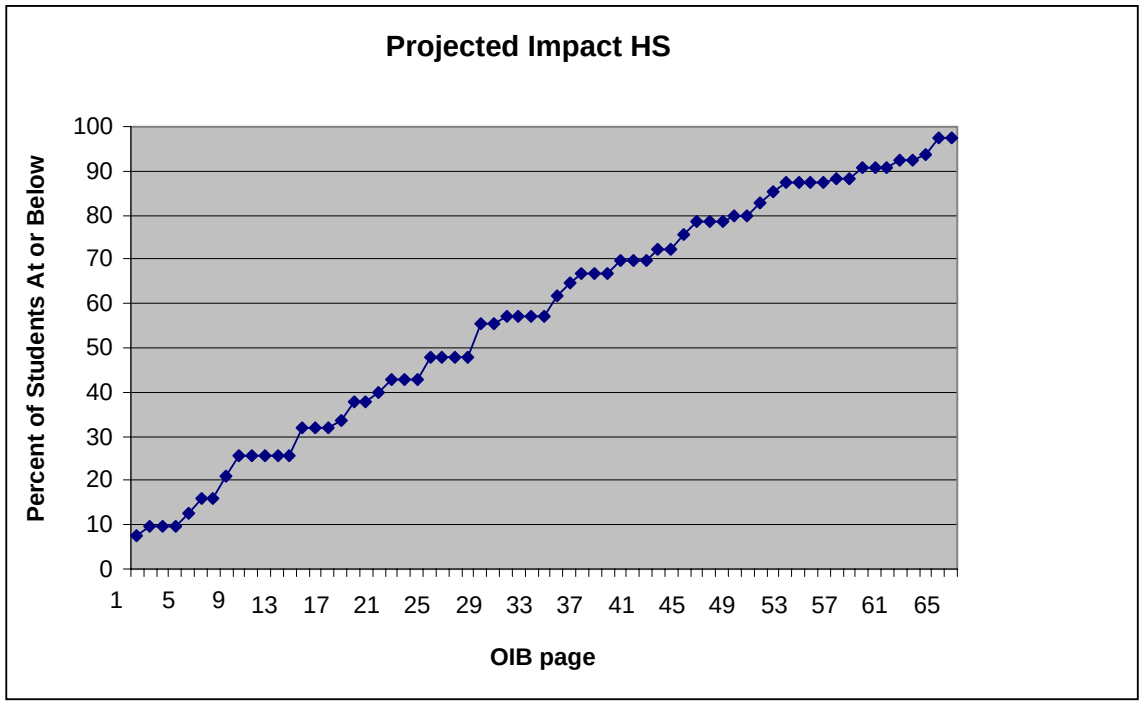


Figure 3. Projected Impact Data using RP67 for HS Form A.



Figures 4–6 address the NAAAC's request to see RP55 compared with RP67.
Because the AIMS test uses the Rasch model, items stay in the same order in the OIB, however, the impact data changes because the score to which an item is mapped changes. As is to be expected, using RP67 makes the tests more difficult in terms of cut score locations. That is, fewer students score at or below each page of the OIB with RP67 than RP55. Also noticeable are the technical issues presented earlier. Those issues do not change it is just the bookmark location of those gaps or item groupings that shift.

Figure 4. Projected Impact Data using RP55 and RP67 for Grade 4 Form A.

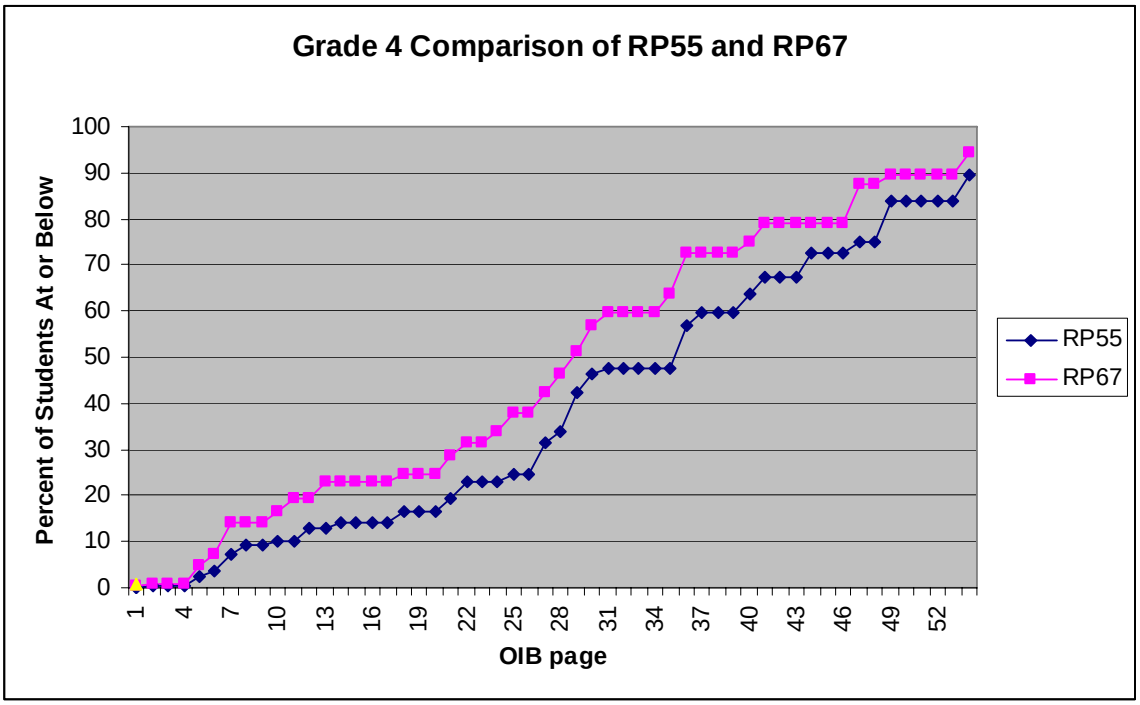


Figure 5. Projected Impact Data using RP55 and RP67 for Grade 8 Form A.

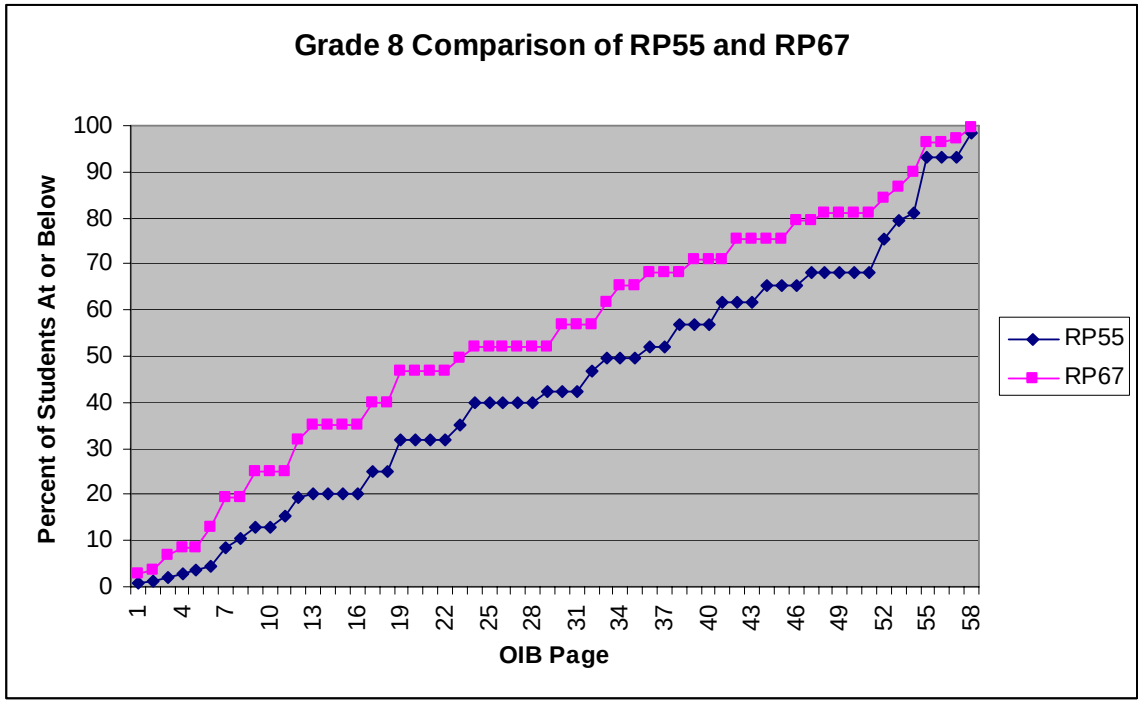
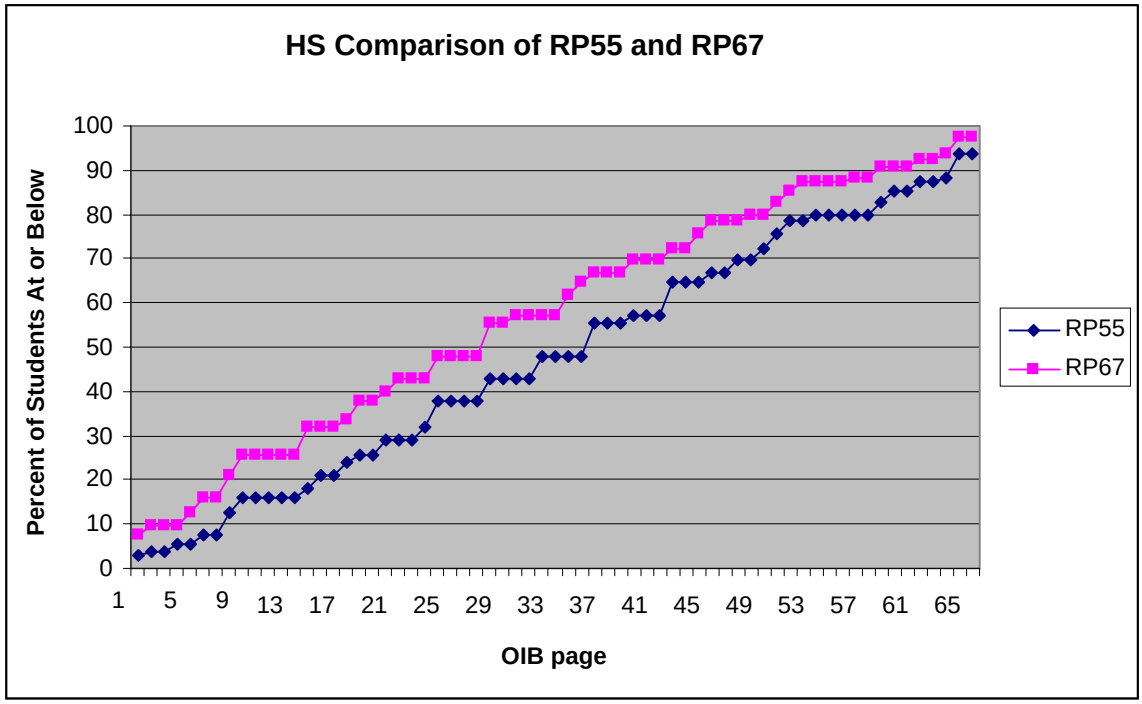


Figure 6. Projected Impact Data using RP55 and RP67 for HS Form A.



In Figures 7–9 RP67 is used to address the originators of the BSSP suggestion that OIB's span from about 80 to 110 score points to exceed normal test length. To accomplish this, both Form A and Form B were combined – anchor items appear only once– to meet the target length of an OIB. While there are still multiple points with items at the same bookmark location, the additional items reduced, in most cases, the magnitude of the gaps is diminished: there are fewer large jumps in impact data based upon a one-page moves of the bookmark in the OIB.

More items along the theta distribution permit more precise cut scores to be recommended, and more items help participants refrain from "percent correct" thinking on one test. Instead, longer OIBs assist participants in generalizing the knowledge, skills, and abilities students possess when they answer particular items correct based upon the content. Moreover, including more items in the OIB is beneficial for participants when writing the performance level descriptors (PLDs).

Figure 7. Projected Impact Data using Form A and B combined: Grade 4.

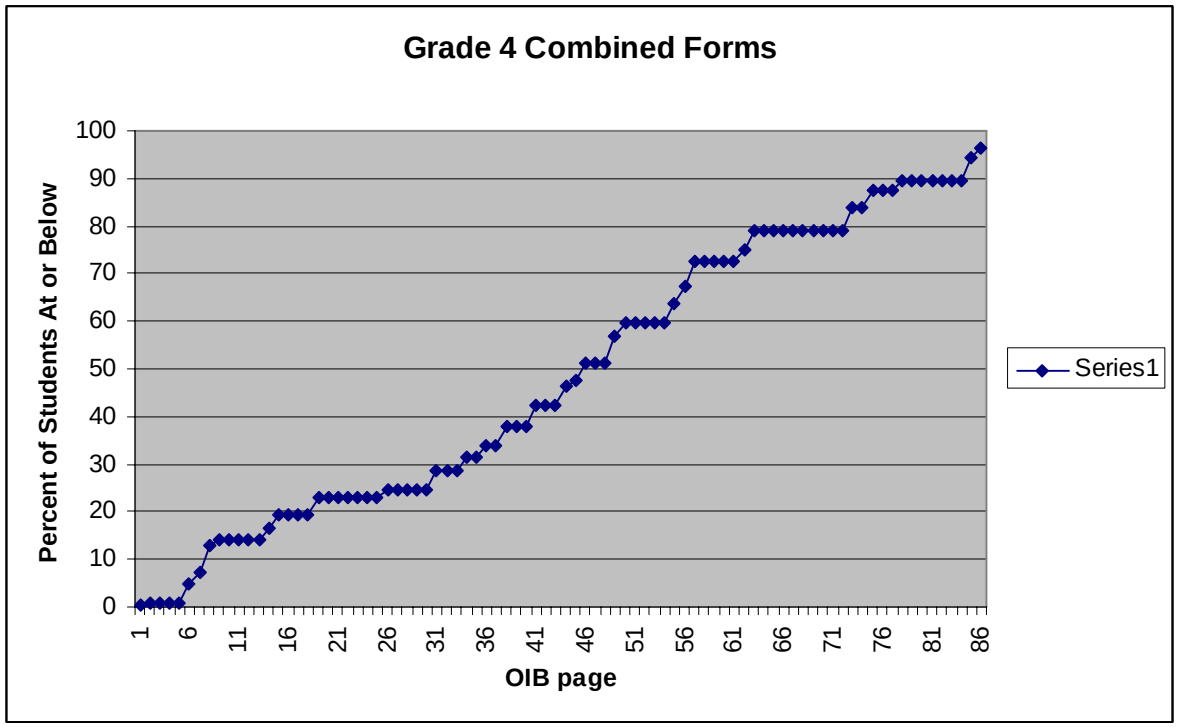


Figure 8. Projected Impact Data using Form A and B combined: Grade 8.

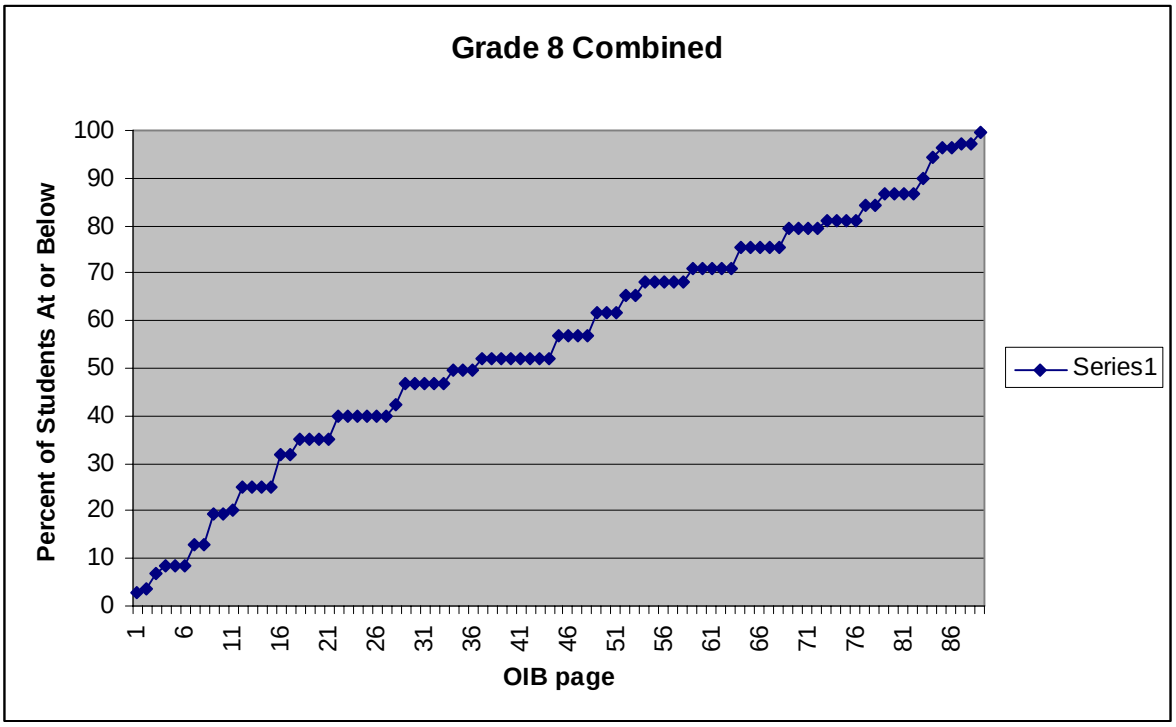
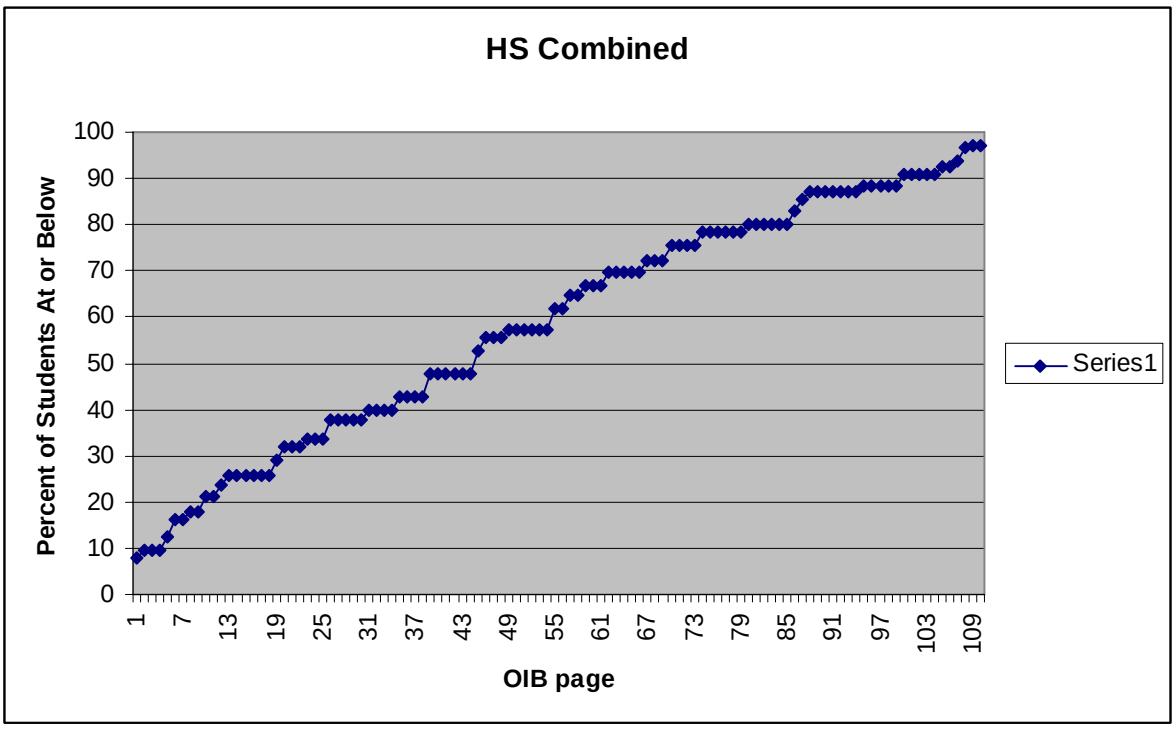


Figure 9. Projected Impact Data using Form A verses Form A and B combined: HS.



Discussion

When making decisions regarding the methodology for creating OIBs, one of the main considerations is whether participants have enough information to make decisions about students in each achievement level. To the degree possible, the items should be evenly spread along the achievement continuum. In these analyses, we explored the distribution of items in order to understand the feasibility of using an IRT approach for the AIMS Science Bookmark standard setting. **The analyses support the use of IRT BSSP**.

The choice of RP value is an important policy decision that must be made before a standard setting that uses an item mapping procedure such as the BSSP. However, in addition to being a policy decision, there are also technical and practical considerations that should be taken into account based upon the items present in the form used for standard setting purposes. The above analyses illustrated the ways different RP values can affect how well the items cover the range of student abilities along the scale, the item ordering of the OIB, recommended cut scores, and percentages of students classified in each performance level.

The projected data suggest it is possible to use an IRT approach to order items; however, the data do not provide a clear choice in terms of the RP criterion for a single intact form. The Grade 4 data suggest the use of RP65 or RP67 while the Grade 8 data suggest the use of RP60 and HS data suggest the use of RP55 or RP60. For all three grades, when Forms A and B are combined, RP67 works well to describe the range of student abilities. Multiple items at the same bookmark location are present when RP67 is used, although this occurs no matter the RP value used or whether forms are used intact or combined.

The use of multiple RP values at a single standard setting is not recommended. Additionally, Zwick, Senturk, Wang, and Loomis (2001) found that in general content experts felt comfortable with RP values around 0.70. Zwick, et al. determined RP values of 0.65 – 0.74 align more closely to content experts' expectations.

Considering that the data in Tables 2-4 which suggest RP values ranging from RP55 to RP67; that RP67 works well when Forms A and B are added together to form the OIB; that literature supports RP values around 0.70 (Zwick et al., 2001; Lewis et al., 1996); and that common standard setting practice uses RP67, **CTB recommends the use of RP67** with the creation of a pseudoform for the standard setting.

The standard setting pseudoform would comprise items from both forms, would represent the blueprint of the test, would describe the ranges of student abilities, and would be longer than a single intact test form. Moreover, the pseudoform would minimize the number of items mapped to the same location of the scale.

These results should be viewed with caution because they are based on student score projections and field test data. The field test forms were shorter than the intact test forms will be, and approximately 2,000 students took each item. Once all items are calibrated in intact forms (i.e., item parameters are calculated on the census population for the forms), it can be expected that the parameters will change. Additionally, it can be expected that the parameters will become

more precise due to the increased information that more students and more items bring to the estimation process. Similarly, student scores will also become more precise. Therefore, it is important that the data be reviewed again with the operational parameters.

Conclusion

The field-test AIMS Science data support the use of an IRT BSSP approach. The data also suggest that the two forms should be used to create a single OIB. To implement a successful standard setting for the AIMS Science, the CTB Standard Setting Team will plans to apply the IRT BSSP approach to AIMS Science pseudoforms, as described above, and will allot appropriate amounts of time in the project schedule to provide for these tasks.

Participants, Materials, & Implementation

During the standard setting meeting, participants will review the content of the AIMS Science assessments using the BSSP. These sections will describe the types of participants that should be recruited to each meeting, the materials used at the BSSP, and the implementation of the BSSP.

Types of Participants

Participants will recommend cut scores in Grades 4, 8, and high school. The standard setting committee for each grade level should represent a sample of expert participants from the entire pool of all such qualified experts. It is important that the sample is representative of the pool of experts in terms of geographic location, socioeconomic status, ethnicity, gender, community size, and other demographic characteristics. Teacher-participants should be proportionally selected from general classrooms, gifted, special education, and vocational classrooms. School/district administrators may also be invited to attend the standard setting as participants.

The department may choose to include parents or other community members in the standard setting. If these types of participants are to be invited, it is important that they be very familiar with and understand the content being tested. Participants who do not understand the content on which cut scores are being established may erode the credibility of the process.

Please see the Part III of the *Bookmark Standard Setting Handbook* (2005) for more information on recruiting participants.

Standard Setting Materials

Two materials are key to the BSSP: the ordered item booklet (OIB) and the item map. This section describes each of these materials.

Ordered Item Booklets

The *ordered item booklets* (OIBs) will comprise multiple-choice (MC) items. These items are ordered in terms of difficulty. The ordering is straightforward in that easier items are placed earlier in the booklet and harder items follow.

Items from Form A and Form B will comprise the items in the OIBs for AIMS Science Grades 4, 8, and high school. The use of items from both forms will provide cut score review participants with a fuller range of the Science content/skills tested.

The *Bookmark Standard Setting Handbook* (2005) found in Appendix A describes the typical, IRT-based implementation of Bookmark.

Item Maps

The *item maps* summarize information about the items in the OIBs. The item map indicates the order of difficulty, scale location, item number, scoring key, and standard that each item measures. On the item map, the participants answer two questions as they examine the items: (1) "What does this item measure? That is, what do you know about a student who can respond successfully to this item/score point?" and (2) "Why is this item more difficult than the preceding items?"

Standard Setting Implementation

The BSSP will bring participants together from across Arizona to set performance standards on the AIMS Science. These participants will be selected by ADE. Participants will be assigned to work in one of the three grades (4, 8, or high school).

For each grade level, there will be approximately 12 participants, including three Table Leaders. The standard setting committee will be divided into grades, each of which will have its own breakout room. For example, Grade 4 Science and Grade 8 Science will meet in different breakout rooms. The implementation of the BSSP will consist of training, orientation, four rounds of judgments, and description writing.

Figure 10 shows the agenda for the AIMS Science standard setting. A description of each day planned for the standard setting follows.

Figure 10. AIMS Science Standard Setting Agenda.

Day 1

8:00 – 8:30	Table Leader Training
8:30 – 9:00	Continental Breakfast
9:00 – 10:30	Opening Session: * Introductions * Housekeeping – Non-Disclosure, Travel, Breakout Room Assignments * Purpose (ADE) * BSSP Process (CTB)
10:30 – 10:45	Break
10:45 – 12:00	Breakout Rooms: * Distribute Performance Level Descriptors * Science Standard and the Target Students Discussion
12:00 – 1:00	Lunch
1:00 – 2:30	Breakout Rooms – Science Pseudo Assessment (Individual Activity)
2:30 – 2:45	Break
2:45 – 3:00	Breakout Rooms – Review Answers to the Science Pseudo Assessment
3:00 – 4:30	Breakout Rooms – Round 1 (Individual Activity)

Day 2

8:30 – 9:00	Continental Breakfast
9:00 – 12:00*	Breakout Rooms: * Review Results of Round 1 (Table Discussion) * Round 2 (Individual Activity)
12:00 – 1:00	Lunch
1:00 – 1:15	Breakout Rooms – Review Impact Data Based on Round 2
1:15 – 4:30*	Breakout Rooms – Round 3 (Grade Level Group Discussion)

**Breaks as determined by group*

Day 3

8:30 – 9:00	Continental Breakfast
9:00 – 10:30	Breakout Rooms: * Review Results of Round 3 (Grade Level Group Discussion)
10:30 – 10:45	Break
10:45 – 12:00	Cut Score Articulation Across Grade Levels
12:00 – 1:00	Lunch
1:00 – 2:30	PLD Revision by Large Grade Level Group
2:30 – 2:45	Break
2:45 – 4:30	PLD Modification Across Grade Levels All participants will complete an evaluation of the BSSP

Day 1

During the training, Table Leaders will receive a brief explanation of their function and the process involved with the Bookmark Standard Setting Procedure. There will be one CTB Group Leader and three Table Leaders per grade level. Group Leaders will facilitate the implementation of the BSSP process and the Table Leaders will have the responsibility to facilitate discussion at their tables, keeping their group focused on the task at hand and finding the middle ground between participants when necessary. An overview of BSSP will be provided by CTB staff.

The Opening Session will provide all participants with information regarding general housekeeping, the purpose for conducting the standard setting, and the BSSP process. Locations will be provided concerning restrooms, breakout rooms, and the room where lunch will be served.

Once participants are in their breakout rooms, their particular grade level's Science Performance Level Descriptors will be distributed. A Target Students discussion will be conducted using the PLDs and the Science Standard.

After lunch, participants will individually complete the Science Pseudo Assessment (combination of Forms A and B). After reviewing the answers to the Science Pseudo Assessment, participants will receive an additional training session. During that training session, participants will again be oriented to the process of setting bookmarks in the ordered item booklet, and how bookmark placements are interpreted. After this training, participants will be given a mid-process evaluation, where they will be tested on their understanding of bookmark placement. CTB will use the results of this mid-process evaluation to gauge participants' understanding of the process, will answer any questions participants may have, and will then instruct participants to set their Round 1 bookmarks.

Round 1 will be performed; whereby, participants will individually study the items, take notes on the item maps, review the Target Students descriptions, and place bookmarks in their ordered item booklets.

Day 2

In the morning of Day 2, upon completion of a table discussion concerning the results of Round 1, participants will perform Round 2. In Round 2, participants will again place their bookmarks independently.

In the afternoon, impact data will be presented to the Grade Level Groups based on the Round 2 median results of all grade level participants. These data will illustrate how cut points selected in Round 2 would impact the student performance level distributions. ADE staff must be in attendance for this presentation. Following the impact data review, Round 3 will be performed through discussion at the tables.

Day 3

Day 3 begins with a review of impact data by the Grade Level Group based on the three tables' median results from Round 3. Cut score articulation across grade levels will then be performed. This will be done in a single room with all participants from the three grade levels present. Cross-grade impact data will be reviewed, and the articulation of data across the grades will be conducted.

After lunch, the final activity for the day concerns the revision of the Science Performance Level Descriptors. Each Large Grade Level Group will meet in their breakout rooms to determine modifications to their grade level's PLDs, and upon completion, the three grade levels will again come together to modify the PLDs across grade levels.

Cut Score Articulation Across Grade Levels

Upon completion of the standard setting, participants will meet to review the articulation of impact data across the grades and to smooth data (if necessary).

Articulation

The articulation of impact data refers to the way these data look across grades. Table 6 shows examples of well-articulated impact data, where the percentage of students classified as *Meets the Standard* and above is constant, increasing, or decreasing across the grades. Poorly-articulated impact data may rise and fall from year to year. The public sometimes expects to see well-articulated impact data because the data meet their expectations for what a test should look like. On the other hand, "poorly-articulated" impact data may reflect the increasing and decreasing expectations and skills expected of students in each grade. Table 7 shows two examples of poorly-articulated impact data. Notice that in Example 1 the percentage of Proficient students is lower in Grade 8 than it is in Grade 4 or high school.

Table 6. Well-articulated impact data, in terms of percentage of students classified as *Meets* or above, by grade. (Note: sample data shown.)

	4	8	HS
Constant	30%	30%	30%
Decreasing	35%	20%	15%
Increasing	20%	35%	40%

Table 7. Poorly-articulated impact data, in terms of percentage of students classified as *Meets* or above, by grade. (Note: sample data shown.)

	4	8	HS
Example 1	30%	19%	35%
Example 2	25%	35%	12%

Adjusting the Cut Scores

Following the presentation of final results, all participants will be convened to examine the impact data associated with their group's recommendations. The purpose of this smoothing discussion is to establish a system of cut scores that is well-articulated and, at the same time, considerate of the participants' original recommendations. Representatives from CTB and ADE will facilitate the cross-grade smoothing discussions.

The ADE wishes to include all participants in the post-standard setting review. By including all participants, each Arizona educator who participated in the standard setting will be able to see how his or her group's recommendations compare with those from other groups, and how the overall system of cut scores works together.

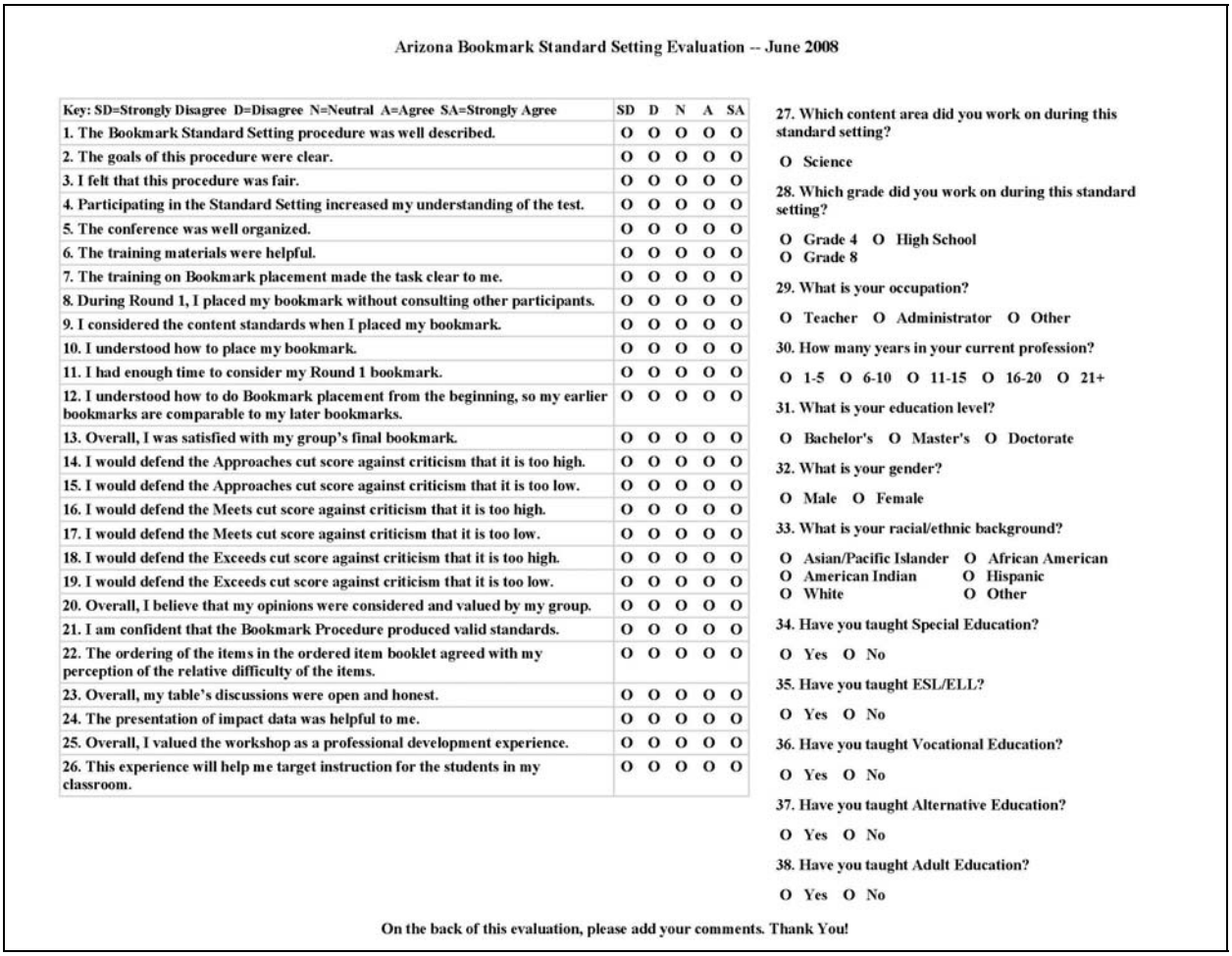
Performance Level Descriptor Revision

Following the Post-Standard Setting Review, participants will engage in a Performance Level Descriptor Revision activity. Throughout the Bookmark Standard Setting Procedure, participants will define and discuss the Target Students. The first Target Student definitions developed by the participants are based on participant expertise, Arizona Content Standards, and Arizona Performance Level Descriptors. Throughout the BSSP, participants will gain new knowledge and insights that help them refine their Target Student definitions. During this time, they may discover that a particular skill was more difficult for students than they originally thought or they may find the opposite to be true. They will refine their Target Student definitions as they work through the standard setting process. The Performance Level Descriptor Revision of the standard setting workshop represents the culmination of these discussions about Target Students.

Evaluation

At the conclusion of the standard setting, participants will be asked to complete a written evaluation of the workshop, as illustrated in Figure 11. Participants will be asked to rate how satisfied they are with the recommended cut scores and with the process. Participants will also be asked to indicate which demographic groups to which they belong.

Figure 11. Sample evaluation of the standard setting.



Arizona Bookmark Standard Setting Evaluation -- June 2008

Key: SD=Strongly Disagree D=Disagree N=Neutral A=Agree SA=Strongly Agree	SD	D	N	A	SA
1. The Bookmark Standard Setting procedure was well described.	O	O	O	O	O
2. The goals of this procedure were clear.	O	O	O	O	O
3. I felt that this procedure was fair.	O	O	O	O	O
4. Participating in the Standard Setting increased my understanding of the test.	O	O	O	O	O
5. The conference was well organized.	O	O	O	O	O
6. The training materials were helpful.	O	O	O	O	O
7. The training on Bookmark placement made the task clear to me.	O	O	O	O	O
8. During Round 1, I placed my bookmark without consulting other participants.	O	O	O	O	O
9. I considered the content standards when I placed my bookmark.	O	O	O	O	O
10. I understood how to place my bookmark.	O	O	O	O	O
11. I had enough time to consider my Round 1 bookmark.	O	O	O	O	O
12. I understood how to do Bookmark placement from the beginning, so my earlier bookmarks are comparable to my later bookmarks.	O	O	O	O	O
13. Overall, I was satisfied with my group's final bookmark.	O	O	O	O	O
14. I would defend the Approaches cut score against criticism that it is too high.	O	O	O	O	O
15. I would defend the Approaches cut score against criticism that it is too low.	O	O	O	O	O
16. I would defend the Meets cut score against criticism that it is too high.	O	O	O	O	O
17. I would defend the Meets cut score against criticism that it is too low.	O	O	O	O	O
18. I would defend the Exceeds cut score against criticism that it is too high.	O	O	O	O	O
19. I would defend the Exceeds cut score against criticism that it is too low.	O	O	O	O	O
20. Overall, I believe that my opinions were considered and valued by my group.	O	O	O	O	O
21. I am confident that the Bookmark Procedure produced valid standards.	O	O	O	O	O
22. The ordering of the items in the ordered item booklet agreed with my perception of the relative difficulty of the items.	O	O	O	O	O
23. Overall, my table's discussions were open and honest.	O	O	O	O	O
24. The presentation of impact data was helpful to me.	O	O	O	O	O
25. Overall, I valued the workshop as a professional development experience.	O	O	O	O	O
26. This experience will help me target instruction for the students in my classroom.	O	O	O	O	O

27. Which content area did you work on during this standard setting?

O Science

28. Which grade did you work on during this standard setting?

O Grade 4 O High School
O Grade 8

29. What is your occupation?

O Teacher O Administrator O Other

30. How many years in your current profession?

O 1-5 O 6-10 O 11-15 O 16-20 O 21+

31. What is your education level?

O Bachelor's O Master's O Doctorate

32. What is your gender?

O Male O Female

33. What is your racial/ethnic background?

O Asian/Pacific Islander O African American
O American Indian O Hispanic
O White O Other

34. Have you taught Special Education?

O Yes O No

35. Have you taught ESL/ELL?

O Yes O No

36. Have you taught Vocational Education?

O Yes O No

37. Have you taught Alternative Education?

O Yes O No

38. Have you taught Adult Education?

O Yes O No

On the back of this evaluation, please add your comments. Thank You!

Technical Report

Within five business days following the standard setting (or by June 18, 2008) CTB will provide a Preliminary Standard Setting Technical Report to the ADE. The Preliminary Report will include the results of the standard setting, the results of the written participant evaluation administered after the standard setting, and information about the standard errors of measurement which surround each recommended cut score.

Within 40 business days following the standard setting (or by August 7, 2008), CTB will provide a Final Standard Setting Technical Report to the ADE. This report will contain detailed information about judgments made by participants in each grade; information about standard errors of measurement and of the bookmark; graphical representations of participants' judgments; detailed summaries of participants' evaluations; and copies of the handouts and overhead slides used during the standard setting workshop. Appendix B contains information on how standard errors of the bookmark (cut score) are calculated.

Appendix A:
See *CTB Bookmark Standard Setting Handbook* (2005)

Appendix B:
Calculating a Meaningful Standard Error for the Bookmark Cut Score

In the Bookmark Standard Setting Procedure for a given grade level, participants are assigned to roughly equivalent small groups that work independently through Round 2. Thus, the set of Round 2 cut scores provide some information about the stability of consensus in Bookmark cut scores across independent small group replications. To quantify this degree of consensus, we calculate the cluster sample standard error (Cochran, 1963, p. 210) of the Round 2 mean cut score. Cluster sample standard errors are appropriate when, as may be reasonably assumed here, data are collected from groups and independence can be assumed between groups but not within groups.

For the Bookmark Procedure, the standard error of the Bookmark cut score (SE_{cut}) is based on the cluster sample standard error of the Round 2 mean cut score. Because the final Bookmark cut scores are based on the *median* of the group instead of the mean, this cluster sample standard error (SE_{cut}) is adjusted by $\sqrt{\frac{\pi}{2}}$ (Huynh, 2003). The standard error of the Bookmark cut score is:

$$SE_{cut} = \left(\sqrt{\frac{\pi}{2}}\right)\left(\sqrt{\frac{S^2}{N}\left[1+\left(\frac{N}{n}-1\right)r\right]}\right),$$

where S^2 is the sample variance of individual Round 2 cut scores, r is the Round 2 intraclass correlation, N is the number of participants, and n is the number of groups. To be precise, if Y_{ik} is the cut score from the i^{th} participant in the k^{th} group, $\overline{Y}_k$ is the average cut score for group k, and $\overline{\overline{Y}}$ is the average of all Round 2 cut scores, then

$$r = \frac{Var(\overline{Y}_k)}{Var(\overline{Y}_k) + Var(Y_{ik} - \overline{Y}_k)} \quad \text{and} \quad S^2 = \frac{1}{N-1}\sum_{n,k}\left(Y_{nk} - \overline{\overline{Y}}\right)^2$$

If we have only two groups ($n=2$) and perfect dependence (agreement) within groups ($r=1$), then the cluster sample standard error simplifies to $SE_{cut} = \left(\sqrt{\frac{\pi}{2}}\right)\left(\frac{|Y_1 - Y_2|}{2}\right)$, which is the standard error formula employed by NAEP for two independent replications of a modified Angoff procedure (ACT, 1983, pp. 4-8). If, on the other hand, individual participants acted independently of their groups ($r=0$), then the cluster sample standard error simplifies to the

traditional standard error of the mean for independent observations, $SE_{cut} = \left(\sqrt{\pi/2}\right)\left(\sqrt{S^2/N}\right)$.

In this manner, SE_{cut} provides a simple, flexible, and general way to quantify the amount of uncertainty associated with final Bookmark cut scores.

It is appropriate (if statistically imprecise) to say that repeated replications of this very standard setting procedure with different judges sampled from the same population of potential judges would result in a range of cut scores, most of which would fall in a band of width 4* SE_{cut}. In the graphical displays of participant data, we depict such an interval centered at the median of the Round 3 cut score. The purpose of calculating statistics like SE_{cut} and producing graphs of the types displayed here is to effectively communicate the complex information that is gathered during a Bookmark Standard Setting Procedure.

References

ACT (1993). Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.

Buckendahl, C.W., Smith, R.W., Impara, J.C., & Plake, B.S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39,* (3), 253-263.

Cochran, W. G. (1963). *Sampling techniques*. New York: John Wiley & Sons.

Huynh, H. (2003, August). Technical Memorandum for Computing Standard Error in Bookmark Standard Setting. (The South Carolina PACT 2003 Standard Setting Support Project). Columbia: University of South Carolina.

Impara, J.C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, (4), p. 353-366.

Kolen, M.J., & Brennan, R.L. (2004). Test equating, scaling, and linking: Methods and Practices. New York: Springer.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, April). *The Bookmark standard setting procedure: Methodology and recent implementations.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). Standard Setting: A Bookmark approach. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment: Phoenix, AZ.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), Setting performance standards (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Zwick, R., Senturk, D., Wang, J., & Loomis, S.C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. Educational Measurement: Issues and Practice, 20, 15-25.